DOCUMENT RESUME

ED 301 167 IR 013 507

AUTHOR Mandeville, Garrett K.
TITLE A Record Linking System for Research on School Improvement.
PUB DATE Apr 88
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algorithms; *Computer Oriented Programs; *Computer Software; Data Analysis; *Database Management Systems; *Data Processing; *Educational Research; Longitudinal Studies; *Student Records
IDENTIFIERS *South Carolina

ABSTRACT

This paper describes and evaluates a computerized system, MATCH, which is being used in South Carolina to match or link student records from year to year. The functioning of MATCH and its specific features are described, and its use in matching student test scores is detailed. The problem of mismatched records and the use of visual editing to lower the rate of mismatches are then discussed. Other current educational research applications of the system are also detailed, including expectancy tables, retention and mobility rates, the evolution of a preschool program, and research on the Program for Effective Teaching. It is suggested that the system provides a way to undertake research and evaluation that has been not been previously possible because of the high costs of creating longitudinal databases for large scale assessment data. The text is supplemented by four tables and an appendix containing rules for deciding whether a match is valid. (7 references) (EW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED301167

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Record Linking System for Research on School Improvement

Garrett K. Mandeville
University of South Carolina

Running head: A RECORD LINKING SYSTEM FOR SCHOOL IMPROVEMENT

Paper presented at the annual convention of the American Educational Research Association, New Orleans, April 1988

IR013507

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Garrett K.Mandevill

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The primary objective of this paper is to describe and present data to validate the effectiveness of a computerized system for matching student test records year-to-year. The system has been used during the past two years to create longitudinally matched student data sets required to apply the methodology used in South Carolina for the purpose of identifying schools to receive monetary awards under the School Incentive Reward Program (SIRP). The validation data presented were obtained during these applications of the algorithm. A secondary objective is to describe a number of school research and evaluation problems which, although very difficult or impossible to attack without the algorithm, have been rendered "doable" with it. These application areas will be enumerated later in the paper.

Validation data of two types will be presented. First, the results from visual editing of approximately 160,000 computer matches for the 1986 and 1987 SIRP analyses will be presented. Second, computer matches will be compared with matches based on numeric ID systems used by two large school districts for the 1987 matched data sets.

Theoretical Framework

The situation in which two data files are available with some records on each file relating to the same person or other object of interest is often referred to as the "record linkage" problem. It is assumed that a valid unique identifier is not present on both files since, if this were the case, the solution would be trivial. Each file, however, must contain data on descriptors such as name and birthdate, some of which may be incorrectly or partially recorded or missing. The objective is to use descriptive information common to both files to correctly link the records.

A number of the studies in this area have involved the medical field.

For example, Phillips, Bahn and Miyasaki (1962) reported on the development of a data base of patient records while Cooley and Cox (1981) discussed a system for matching reports of recipients of medical care with those of providers. In another study, Freeman (1983) reported on a project to link Census of Agriculture records for 1976 and 1981.

Recently, Jaro (1984) and Kelley (1984) noted the advantages of what is called "blocking" the records to be matched. Briefly, blocking involves grouping the records on each file into mutually exclusive and exhaustive categories and restricting the search for matches to records which occur in the same blocks.

A Brief Description of the Matching Algorithm

It would clearly be inappropriate in this paper to attempt to describe in any detail how the matching system (MATCH) operates. However, the following brief description of the main features does seem appropriate.

Blocking Variables

MATCH uses gender, ethnicity, birth month (exclusive of year) and school district as blocking variables. Ethnicity codes other than black and white are combined into a third "other" category, to be taken in combination with the two gender and 12 birth month categories to produce 72 blocks within each of the 92 school districts, yielding a total of 6624 blocks statewide. An obvious advantage in grouping records into blocks is to reduce the number of comparisons in attempts to identify matches. For example, in South Carolina approximately 50,000 students are tested at most grade levels each year. If each of two files containing 50,000 records (e.g., Spring 1986 1st grade records and Spring 1987 2nd grade records) were matched by comparing each record from one file with each record from the other file, a total of 2.5 billion comparisons would be required. This approach would be inefficient and would strain even a modern mainframe

computer.

A more efficient strategy is to use the selected blocking variables to group the records on both files into the 6624 mutually exclusive and exhaustive categories and restrict the inter-file comparisons to records in the same categories. For example, suppose that the 50,00 records were evenly distributed among the 6624 categories, yielding about 8 records per category. Then only 64 comparisons would be necessary in each of the 6624 categories for a total of only about 420,000. Although this is still a large amount of data processing, it is obviously more efficient in that only about .02 percent as many comparisons are necessary. The fact that the records are not evenly distributed across the categories, while invalidating the numbers above, does not invalidate the principle; blocking does render an impossible problem manageable. (It should be noted that for the SIRP application, the restriction that the child remains in the same school district for the two years is reasonable; for other applications it may be removed.)

MATCH Operates in Passes

The second general strategy which was incorporated into the MATCH is that it operates in discrete steps or "passes." During the 1st pass, all descriptors (including birthdate and name) on the two files must be identical, except that the two names may differ in a certain number of characters (this parameter is usually set to one). Generally, between 60 and 70 percent of the records are matched on this 1st pass. The main factor which affects this matching rate is the comparability of the name fields from the two files. (For example, on one set of test records a single 15 character name field is used whereas the more common format provides separate fields for first and last name and middle initial which total 28 bytes.)

During the 2nd pass, the residual records (those not matched on the 1st pass) are processed along lines similar to those of pass 1, except that the two names need not be as consistent as required in pass 1. During this processing step, each file 1 record is compared to all file 2 records in the same category. A number of heuristic rules concerning the likelihood that the names on the two files represent the same child are applied and a "pseudo-probability" of this event is computed. The rules which have been operationalized in the algorithm were based on inspecting a large number of records and inferring the types of inconsistencies which arise. (It should be noted, however, that children invent new ways to disguise their names each year so that MATCH is annually revised in an attempt to more adequately simulate human judgement.)

During the 3rd processing step, matches are considered if the records are consistent on school district and on two of the other three descriptors, i.e., gender, ethnicity and birthday. In an attempt to reduce the number of mismatches, however, the "pseudo-probability" required for a match to be considered valid is increased.

Other Features of MATCH

The above discussion has concentrated on the record-linking capability which is the heart of the system. The matching is accomplished by a computer program of approximately 1200 statements written in the FORTRAN 77 language. MATCH also includes pre- and post-processing job steps which utilize the Statistical Analysis System (SAS Institute, 1985). During pre-processing one of the files, file #1 say, is read and a SAS data set is created for the variables of interest. As a second pre-processing activity, two additional SAS data sets are created (one for each file) which contain information on the matching variables only. This latter information is passed to the FORTRAN program which identifies the matches

and writes out the desired information from the second file, file #2 say, along with the linking identifier needed to associate these records with the correct file #1 record. This information is then passed to the SAS post-processing step during which the file #2 information is read into a SAS data set and merged with file #1.

A convenient feature of MATCH is that the user may identify the specific fields of file #2 for which data are to be retained. This is important because the statewide data files generally include item level data which are not of interest but which cause these records to be lengthy (often exceeding 1000 bytes). To enhance the flexibility of the system, the variable names the user desires, their field specifications on file #2, and type (numeric or character) are input to the FORTRAN program which handles all data transfer activities and passes an appropriate INPUT statement to the SAS post-processing step.

Methods and Data Sources

As noted above, MATCH has been used for the past two years to create data sets for the South Carolina SIRP. In 1980 the South Carolina General Assembly passed legislation mandating the development of statewide objectives in reading and mathematics and the annual statewide assessment of student achievement of these objectives. As a consequence the Basic Skills Assessment Program (BSAP) was implemented with criterion referenced tests administered in the spring of each year at grades 1, 2, 3, 6, 8, and 10. These tests are relatively short, typically containing 36 reading items and 30 math items, six for each of the six reading and five mathematics objectives. Writing is assessed with a writing sample and only at the higher grades (i.e., 6,8 and 10). (Note: the 10th grade test has recently been redesigned to serve as a high school Exit Exam and is somewhat longer than the other BSAP tests.)

In addition to the BSAP, two other statewide testing programs are in operation. At the beginning of the school year the Comprehensive Skills Assessment Battery (CSAB) is individually administered to each 1st grader as a school readiness test. Furthermore, various levels of the Comprehensive Test of Basic Skills (CTBS) are given to all students who have completed grades 4, 5, 7, 9 and 11. Under all of these test programs, information on special programs (e.g., handicapped, gifted, Chapter I) in which the student may be enrolled, student descriptors (e.g., names, birthdates, gender, and ethnicity) and a school identifier are coded onto the machine readable answer sheets in addition to the students' responses. In the lower grades, teachers complete all information except the test answers.

In the Spring of 1986, the Exit Exam test was under development so no testing took place at the 10th grade. For this reason the 1986 SIRP was based on test data for grades 1-9 and 11. In 1987 the Exit Exam was utilized so the SIRP for that year was based on test data for all grades in the 1-11 span.

SIRP Matching for 1986 and 1987

The SIRP analytic model employed in South Carolina utilizes student level regression analyses in which the most recent (posttest) scores are regressed on the previous years (pretest) scores. The most recent spring test data serve as the posttest and usually the previous spring scores serve as the pretest. The single exception is that for the 1st grade, the previous fall CSAB records constitute the pretest data.

The first step in the data processing was to match pretest and posttest records for students who were either in the 1st grade or who were in higher grades and were not retained in the same grade during the current year. This is referred to as Initial Match Processing (IMP). Using the 1987 SIRP

as an example, the BSAP1 records for Spring 1987 were matched against the CSAB records from Fall of 1986, the BSAP2 records for Spring 1987 were matched with the BSAP1 records for Spring 1986, etc. Typically 80% to 90% of the posttest records for any grade level are matched during IMP.

At the conclusion of IMP, a complete listing of pass 2 and pass 3 matches and a listing of the unmatched records from the two files was printed and visually edited. The editing had two objectives and was conducted according to a set of rules as to what would constitute a "valid" match (the rules are in an Appendix). One objective was to scan the computer matches from passes 2 and 3 to identify probable mismatches (referred to as DELS for delete) so that they would not be included in the analysis. The other objective was to scan the two residual files (the unmatched records) to identify what appear to be valid matches that the computer missed (referred to as JOINS) so that they could be correctly linked and used in the analysis. The results of this verification process will constitute one type of evidence concerning the effectiveness of the computer matching algorithm.

The objective of the next step in the SIRP analysis was to match the test records of students who were retained in grades 2-11 since they were not matched during IMP. (Since 1st grade repeaters are retested on the CSAB and BSAP1 they are identified during IMP.) This was accomplished by matching the proper "residual" files from IMP. In order to clarify this processing, it will be necessary to describe the situation when IMP has been completed.

At the conclusion of an IMP run for a given grade level, three output files are created. For example, for the 1987 analysis at grade 3, the output files (also SAS data sets) were labeled G23_8687, F23_8687, and L23_8687. The "G" (good) file contained the computer matches, the "F"

(first) file the nonmatched 2nd grade 1986 records and the "L" (last) file the nonmatched 3rd grade 1987 records. Similar files were created at each grade level.

Consider a student who was retained in grade 2 for both the 1985-86 and 1986-87 school years and, therefore, was tested with the BSAP2 in the Spring of 1986 and the Spring of 1987. This student's pretest record should be on the F23_8687 file (2nd grade in 1986) and his or her posttest record should be on the L12_8687 file (2nd grade in 1987). This stage in the SIRP analysis was designed to link the data for these retained students and is referred to as Repeater Match Processing (RMP). Visual editing of these matches was limited to DEL editing only.

During the collection of the 1987 test data, two large school districts implemented student tracking systems based on identification numbers (IDs). These numbers were coded into a portion of the answer sheet called the "special codes" field. During the matching for the 1987 SIRP analysis these codes were "captured" along with the test and other information which is routinely included in the resultant data files. A comparison of the results using the ID numbers and those produced by the matching algorithm constitutes the second form of evidence reported below.

Differences in 1986 and 1987 Matching

It should be noted that, in an effort to upgrade the process, there were some differences in the way the matching was conducted during the two years. For example, since RMP utilizes the F and L files, these two files should be rewritten to reflect the changes implied by the visual editing before RMP processing is conducted. In order to accomplish this the DELS from the G files would be "split" and included in the new F and L files and F and L file records used to create the JOINS would be excluded. This extra processing was done in the 1987 analysis but, for lack of time, not

during 1986. Also ID numbers (for the two districts) were used in 1987 to supplement JOIN/DEL editing and this was not true in 1986.

Another difference was that the program parameters were "tightened" for the 1987 SIRP computer runs so that they more adequately reflected the rules regarding what constituted a valid match. This occurred because the rules were not in place at the beginning of the 1986 analyses but evolved during processing. Thus we might expect a somewhat lower mismatch rate (fewer DELS) for 1987 than 1986 but probably a larger rate of missed matches (JOINS).

Results

Visual Editing

Over 800,000 records were matched by the computer during IMP for the two years with about 160,000 (20%) subject to visual editing (pass 2 or 3) and the data on DEL and JOIN edits are presented in Table 1.

Table 1 about here

For 1986, visual editing revealed 1736 DEL records which were designated as possible mismatches out of the total of 383,725 for a mismatch rate of 0.45%. For 1987, with the more stringent definition as to what was required for a valid match, only 972 of the 427,056 records were designated as DELS so that the mismatch rate was roughly cut in half (0.23%). As far as missed matches are concerned, in 1986 1123 JOINS were found for a rate of 0.29% and in 1987 the figure increased to 2887 or 0.68%. It is worth noting that the tightening of the program parameters lead to a larger number of JOINS for two reasons. The first and obvious one is that more legitimate matches will be missed. The second less obvious reason is that current visual editing procedures do not provide for

reJOINing records which are identified as DELS during visual editing.

MATCH vs. ID Numbers for 1987 Data

The two districts which implemented the numeric ID systems are large by South Carolina standards containing between 2000 and 4000 students per grade level. For the initial analysis, matches based on the IDs were created and compared with those identified by MATCH. One (the most simple) kind of comparison of the relative effectiveness of the two approaches is to consider the total number of matches produced prior to any editing. The summary results, aggregated across grades, are presented in Table 2.

Table 2 about here

Because these results are unedited, they should be interpreted with caution. They provide clear evidence, however, that unedited numeric IDs are not the panacea they might be expected to be; actually, the computer matching program produced 4% to 5% more matches than the IDs. Although this may not seem like a very large difference, in District A this amounted to a total of 1264 more student records across the 11 grades or about 115 more children matched per grade. For District B the advantage of MATCH over the ID number system was even greater with 2198 more records matched or 200 more per grade. On a cost-benefit basis, the matching program, which uses only extant demographic information, has even a larger advantage over the ID system.

There appear to be two reasons why the numeric IDs did not work particularly well. The first is that many were not coded sufficiently to be accurately read by the scanner. For example, on many records one or more blanks were embedded in the "number," rendering it unreadable by the computer. Also, a number of ID fields were completely blank. A second

reason is that there were some duplicate IDs--which renders them inappropriate for merging data sets--often for what appear to be siblings. Thus, if the ID number system were to be effectively used, a considerable amount of time and effort would need to be expended on editing.

The information presented above, although strongly suggesting that the computer algorithm is more effective than using ID numbers for linking student records, is incomplete. The next task was to look more carefully at the records which were inconsistently categorized by the two approaches and estimate "true" error rates. In this analysis, it was assumed that those records on which the two systems were in agreement had been correctly matched.

Inconsistently Linked Records

Printouts were produced for computer matches which were not matched consistently by the two systems. Those linked by MATCH but not by the IDs were reviewed with the objective of identifying additional DELS which had not been noted during visual editing. The primary focus was on pass 3 matches since they are the most unreliable.

Matches implied by the ID numbers but not by MATCH were also listed. Basically, there are two types of such records. They are: 1) combinations of record pairs neither of which were linked (to any records) by MATCH, and 2) combinations of records for which one or both were erroneously involved in other MATCH matches. Valid ID matches of the first type simply represent missed matches and they will be referred to as "pure misses." (Some of these pure misses may also have been identified as JOINs during visual editing). Legitimate linkages in the second category, however, indicate both mismatches created by MATCH and missed matches. For ID matches of this type in which only one of the two records was matched by MATCH one mismatch and one missed match were created; for records in which

both records were inappropriately linked by MATCH, two mismatches and two missed matches resulted. Matches based on the IDs in either category, which were identified as being incorrect, provided information to estimate the mismatch rate for the ID number system.

The primary objective of these analyses was to estimate two error rates for MATCH: the mismatch rate and the missed match rate. Furthermore, each of these error rates was estimated both before and after visual editing. Thus, it was necessary to integrate the DEL/JOIN edit information and the results based on the IDs for the two districts. Since a considerable amount of work was involved, this task was limited to a sample of six of the 11 grades involved, namely grades 1, 3, 5, 7, 9, and 11.

The estimate of the mismatch rate for a given grade and district was found by adding the number of mismatches identified during visual editing to those detected based on the ID linkage information, taking account of any overlap which occurred. The union of these two sets was taken to be the number of unedited mismatches and the unedited MATCH mismatch error rate was estimated by expressing this count as a percentage of the total number of records linked by MATCH. The edited MATCH mismatch error rate was computed in the same way but any mismatches detected as DELs during visual editing were not included. The unedited mismatch rate for the IDs was obtained in a similar fashion, i.e., expressing the number of mismatches as a percentage of the number of records linked based on the ID number system. The results are presented in Table 3 below.

Table 3 about here

We observe that the unedited and edited mismatch rates for MATCH, although larger than the rates based on the ID number system, are very

small indeed. Without editing, about 26 out of 10,000 record linkages produced by MATCH are probably incorrect; after editing, this figure is reduced to about 18 in 10,000. For applications which do not involve decisions about individual students, this is surely a tolerable error rate. With the exception of District A results for 9th graders, the rates are quite consistent across grades and districts. It is also worth noting that this is likely to be an overestimate of the statewide mismatch rate because of the fact that the districts involved are two of the largest in South Carolina. The reason that this is probably an overestimate of the statewide mismatch rate is that for larger districts, the probability that children in the same sex, ethnicity and birthdate categories will have similar names is larger than for smaller districts.

To estimate the rate of missed matches, the first task was to determine the number of correct matches which could have been identified. Of course this cannot be done with certainty, but, considering the collateral ID information, the estimates should be quite good. The initial step in this task was to form the union of the JOINs detected during visual editing and the pure misses based on the IDs. This figure was then supplemented by the number of missed matches which were implied by mismatches. Once again unedited and edited error rates were computed for MATCH and the estimated number of potential correct matches was used in the denominator of these computations. The estimated number of correct matches served as the denominators in these error rate calculations, i.e., mismatches were excluded. The results are displayed in Table 4 below.

Table 4 about here

MATCH without editing misses about 8 in 1000 matches on the average;

with editing, the figure drops to about 5 in 1000 cases. The unedited results are not as consistent (across the 12 samples) as was true of the mismatch rates; the somewhat higher rates for the 1st grade are probably due to the shortened name field associated with the CSAB answer sheets. After visual editing the results are more stable with all but three of the 12 rates in the .30% to .60% range. Again, for most applications, MATCH would seem to be sufficiently accurate.

The use of unedited IDs, on the other hand, may be expected to miss about 6% of the potential matches on the average. The error rate seems to be too large for almost any applications and suggests that any use of the IDs would necessarily require that they first be edited.

One further analysis was conducted on the pure misses in an attempt to better characterize the properties of MATCH. There were 305 such cases and they were categorized as to what appeared to be the primary reason that they were not linked by MATCH. It was also noted as to whether each of these records was identified as a JOIN during visual editing.

A category of particular interest were those due to surname changes. There were 68 such cases or 22.3% of the pure misses and, not surprisingly, none were detected during visual editing. The next largest group were those missed because of coding errors on one or more of the descriptors--sex, ethnicity, and birthday. There were 54 such records, 17.7% of the total, and roughly two-thirds (68.5%) were correctly linked during editing. The use of knicknames such as TONY (for ANTHONY), JUNIOR, SCOTTY, etc., was also a fairly common problem there being 44 such records. This constitutes 14.4% of the pure misses and 81.8% of them were correctly identified during visual editing.

The remaining pure misses were due to a variety of problems with the name, the components of which will be denoted and LAST and FIRST. Included

were embedded blanks in either LAST or FIRST, no blank between LAST and FIRST (CSAB only, since the fields are separate on the other answer sheets), no FIRST coded, 1st letter of LAST missing or miscoded, LAST/FIRST reversals, and leaving one blank space between each character (e.g., J O H N S ON). In a number of cases, multiple coding errors were involved. For example, LAST/FIRST reversals can be detected by MATCH but only in pass 2 during which the descriptors must be in perfect agreement. Thus, valid matches with LAST/FIRST reversals (on one record) and an error on at least one of the descriptors (for either of the two records) will almost surely not be correctly linked by MATCH. This general category contained the remaining 139 pure misses (45.6%) and 56.8% of them were correctly matched during editing. Overall, 49.8% of the 305 pure misses were detected as JOINS during visual editing; if the 68 surname changes are eliminated, 64.1% of the 237 correct matches were found.

Since it is virtually impossible for MATCH to properly link records were the child's surname has been changed, it is useful to have an estimate of the relative frequency of such occurrences. Using the total of the "Number Possible" column of Table 4 (35371) as the denominator, the result is that .19% of the children would be virtually unmatchable. Of course, in small districts it might be possible to infer these JOINs during visual editing but such a practice is quite chancy. (sp??)

Other Applications

Although the context of the discussion above has focussed on the use of the computer algorithm to link student data for the South Carolina SIRP, MATCH or variations of it have been or will in the near furure be used for a number of other applications in support of educational research. These and others which come to mind will now be briefly described.

Expectancy Tables

As noted earlier, South Carolina has recently developed a new Exit Examination which must be passed by high school students (beginning in the 1989-90 school year) if they are to receive a standard diploma. The SCDOE is using MATCH to link 8th grade BSAP and 9th grade CTBS scores to Exit Exam scores so that the two former sets of scores may be used as predictors in the creation of expectancy tables to predict failure on the Exit Exam.

Retention and Mobility Rates

As a natural byproduct, the RMP processing described above provides reasonable estimates of student mobility and retention rates (at the school or district level) which may promote research on the relationships of these variables to student achievement. Although such data are usually gathered by SEAs and LEAs, the validity of the results of these efforts is often questionable.

The Evaluation of a Preschool Program

A modification of MATCH was used to capture test data on the CSAB and BSAP tests for disadvantaged students who participated in a preschool program during the 1984-85 school year. Of the 33 children who entered the 1st grade in the Fall of 1985, 32 were correctly matched on the CSAB and it was eventually determined that the other child had moved out-of-state. On the BSAP1 test given in the Spring of 1986, 31 of the remaining 32 students were correctly matched. During the most recent year (Spring 1987), BSAP scores for all 32 of the original 33 children who still lived in South Carolina were also captured. For the 22 children who were promoted to grade 2, BSAP2 scores were obtained and BSAP1 scores were captured for the 10 who were retained in or returned to the 1st grade. (It should be noted that the use of the modified version of MATCH was not as automatic as in the SIRP analyses; some of the searches for "missing" children involved

more "brute force" methods.)

These results strongly suggest that MATCH may facilitate research which depends upon the creation of longitudinal data sets such as the long term effects of intervention programs on poor children, at risk children, and handicapped children. Even though there have been a number of studies on the issue, there is still doubt as to the effectiveness of early intervention (see, e.g., Clark-Stewart & Fein, 1983, for a review). One reason for the inconclusive nature of these findings is that researchers concerned with this issue have traditionally been hamstrung by the difficulty of, and cost involved in, the conduct of longitudinal studies. For more on this application see Mandeville, Raymond and Anderson (in press).

Research on the Effects of PET

This writer is currently conducting a study to assess the effects of the Program for Effective Teaching (PET) which was implemented in South Carolina on a small scale in 1984 (only a few hundred teachers participated) and has subsequently gained in popularity so that roughly 8400 teachers statewide had had some PET training by the beginning of the 1986-87 school year. Since various (nonrandom) forces operated to determine participation, a "clean" study was impossible. A unique feature of this application is that records with the names of the participating teachers records were matched with other data sets collected by the SCDOE. Matching the PET teachers with one extant data set (the so-called Basic Educational Data System or BEDS data set) was necessary to obtain information on the school the teacher was teaching in, the grade level, the number of students in his/her class, and whether the teacher's class was self-contained. This information was further merged with data on the socio-economic-status (SES) of the school population. This allowed for the

selection of samples of trained and untrained teachers who were teaching in similar school settings.

Next, the classroom headers (which contain the teacher's name) from the statewide testing data tapes were matched with the other records. Eventually, two years of longitudinally matched student test data were attributed to each teacher. The prior years scores served as a partial control of student entry (to the teachers' class) characteristics and the recent scores served as the dependent variables. Although this was a tremendous amount of work and involved many matching runs (and much visual editing) it appeared to be the most rigorous approach to the evaluation of PET. (It should be noted that dummy values of the usual descriptors had to be used in this application since gender, ethnicity, and birthday are not available on all of the teacher data sources.)

Additional Studies to be Conducted by the SCDOE

The SCDOE is currently under contract to use the algorithm for two other studies. The first is the EIA Early Childhood Project which is similar to the one described above. In this study, the 4 year old project participants will be tracked through the 3rd grade with the objective of obtaining the CSAB and BSAP test data and the other information collected during statewide testing.

The second is the Longitudinal Remedial/Compensatory Project. As the name implies, the goal of the study is to monitor the performance of students who are participants in two state mandated programs targeted for low achieving students (the most needy students are served by the Remedial Program and other low-scoring children fall under the Compensatory Program). The primary focus of this study is to study patterns of recividism, i.e., the extent to which students are consistently involved in these programs, "permanently" manage to achieve at levels so that they are

no longer eligible, or alternate between eligibility and noneligibility.

Educational Significance

Educators seem to be much more willing to place credence in the School Incentive Reward Program results based on matched data sets than they were during the first year the SIRP was in operation when school awards were based on year-to-year trends for the (primarily different) students at each grade level. Furthermore, the excuse that a school has a high student mobility rate is partially ameliorated, at least in the sense that the students used in the analysis were (very likely) enrolled in the district for the full year.

Beyond the SIRP, the applications in the previous section speak strongly for the significance of MATCH as a tool to assist in dealing with a number of important educational issues which have either not received much attention or, if they have, only at great expenditure of time and resources. The data on mismatch and missed match rates should be sufficient for most applications. In some cases, visual editing may not be required; in other more demanding applications, more thorough editing may be necessary. In either case, the initial application of the matching algorithm should move the user close enough to a solution so that the remaining "clean up" work is feasible.

The alternative approach of using a system based on the use of edited ID numbers, with the additional costs and inconvenience to students and teachers, should not be necessary for most applications. However, if 100% accuracy is required, a system which can link records for children who have changed surnames would be necessary. If such children were "flagged" in some way during data collection, it would be possible to redesign MATCH to facilitate this. As we have seen above, the IDs used in this study were not 100% accurate. One of the more common problems was that the IDs of

siblings, especially twins, were swapped. It is doubtful that it is possible to develop any system which is 100% accurate for both mismatches and missed matches.

In summary, the MATCH algorithm has the promise of opening up research and evaluation doors which have heretofore been effectively closed because of the tremendous costs involved in creating longitudinal data bases for large scale assessment data.

References

Cooley, P. and B. Cox (1981). An automated procedure for matching record check and household-reported health care data. Proceedings of the Social Statistics Section, American Statistical Association, 418-423.

Freeman, R. W. (1983). 1981 to 1976 census of agriculture record linkage. Proceedings of the Bureau of Economic Statistics Section, American Statistical Association, 91-99.

Jaro, M. (1984). Record linkage research and the calibration of record linkage algorithms. Proceedings of the Social Statistics Section, American Statistical Association, 599-601.

Kelley, R. P. (1984). Blocking considerations for record linkage under conditions of uncertainty. Proceedings of the Social Statistics Section, American Statistical Association, 602-609.

Mandeville, G. K., Raymond and Anderson (in press). The DAISEY data capture system: A computerized system to support longitudinal research. Early Child Development and Care.

Phillips, W., A. K. Bahn and M. Miyasaki (1962). Person-matching by electronic methods. Communications of the Association of Computing Machinery, 5, 404-407.

SAS user's guide: Basics (Version 5 Edition). (1985). Cary, NC: The SAS Institute.

Table 1
Matching Results For IMP Processing By Grade
For 1986 and 1987

Grade	# Post Records	IMP N	IMP Pct	Edits DELS(-)	Edits JOIN(+)	IMP Total N	IMP Total Pct
1986							
1	51841	48161	92.9	299	223	48085	92.8
2	44757	37288	83.3	161	176	37303	83.3
3	44699	37857	84.7	194	127	37790	84.5
4	42141	35825	85.0	192	136	35769	84.9
5	42241	36107	85.5	221	59	35945	85.1
6	43922	37124	84.5	99	49	37074	84.8
7	46639	38496	82.5	115	83	38464	82.5
8	47250	39882	84.4	177	76	39781	84.2
9	50933	41257	81.0	168	125	41214	80.9
11	35882	31729	88.4	110	69	31688	88.3
1987							
1	53845	49828	92.5	39	549	50338	93.5
2	46766	38982	83.4	169	272	39085	83.6
3	44846	37803	84.3	132	232	37903	84.5
4	44646	38377	86.0	44	241	38574	86.4
5	42562	36502	85.8	53	229	36678	86.2
6	43397	36522	84.2	125	298	36695	84.6
7	45947	37952	82.6	59	248	38141	83.0
8	45547	38474	84.5	112	247	38609	84.8
9	49920	39788	79.7	59	192	39921	80.0
10	60683	38825	64.0	143	246	38928	64.1
11	38602	34003	88.1	37	133	34099	88.3

Table 2
A Comparison of Unedited Matching Results
By Method of Matching and District
For 1987 SIRP Data

District	Ns by Method IDs	Ns by Method MATCH	Pct. by Method IDs	Pct. by Method MATCH	Total Post-test Records
District A	26059	27323	74.8	78.4	34852
District B	33334	35532	79.0	84.3	42130

Table 3
Estimated Frequency of Mismatches and Mismatch Rates
For Two Systems of Linking Records

		MATCH			IDs	
Grade	Dist	Number Linked	Unedited Freq(%)	Edited Freq(%)	Number Linked	Freq(%)
1	A	4026	8(.20)	6(.15)	3902	1(.03)
	B	4234	13(.31)	6(.14)	3887	1(.03)
3	A	2658	11(.41)	6(.23)	2505	2(.08)
	B	3071	8(.26)	8(.26)	2867	2(.07)
5	A	2419	5(.21)	3(.12)	2361	1(.04)
	B	3017	6(.20)	3(.10)	2838	0(.00)
7	A	2297	7(.30)	6(.26)	2185	5(.23)
	B	3224	8(.25)	6(.19)	3040	1(.03)
9	A	2168	18(.83)	14(.65)	2077	6(.29)
	B	3243	10(.31)	8(.25)	3032	2(.07)
11	A	1861	4(.21)	3(.16)	1711	4(.23)
	B	2914	5(.17)	5(.17)	2802	0(.00)
Median			.26	.18		.06

Table 4
Estimated Frequency and Rates of Missed Matches
For Two Systems of Linking Records

			MATCH		IDs
Grade	Dist	Number Possible	Unedited Freq(%)	Edited Freq(%)	Freq(%)
1	A	4064	49(1.21)	23(.57)	162(3.99)
	B	4301	80(1.86)	37(.86)	414(9.62)
3	A	2677	30(1.12)	23(.86)	174(6.49)
	B	3087	24(0.78)	13(.42)	220(7.13)
5	A	2433	19(0.78)	12(.49)	72(2.96)
	B	3049	38(1.25)	21(.69)	211(6.92)
7	A	2303	13(0.56)	7(.30)	118(5.12)
	B	3242	26(0.80)	10(.31)	202(6.23)
9	A	2165	15(0.69)	12(.55)	88(4.06)
	B	3261	28(0.86)	15(.46)	229(7.02)
11	A	1861	4(0.21)	1(.05)	150(8.06)
	B	2928	19(0.65)	12(.41)	126(4.30)
Median			.79	.44	6.36

Appendix

Rules for Determining Whether a Match is Valid

1. A letter or letters may have been dropped or added at the beginning or end of the name. A letter may have been omitted or added within the name. Two letters may have been reversed within the name. This type of error is acceptable and will usually be obviously a misspelling. If so many letters are omitted that the name cannot be inferred from one entry, it should not be matched.

2. A correctly matched last name with blank first name and middle initial may not be matched. However, a blank first name on one record is allowed if the middle initials are consistent on the two records provided that the descriptors are consistent and that the last name is not a very common one.

3. A reversal of first name and middle initial is acceptable. (Jones Fred A = Jones Anthony F)

4. Common nicknames may be inferred. (Jones Fred A = Jones Tony F)

5. Some nicknames may be matched with any name of approporiate sex. (Buddy, Trey, Junior, Chip, Sissy, Missy, etc.)

6. There should be no conflict in the first name and middle initial, allowing for reversals and nicknames. (Jones Sara M is not matched with Jones Sara R; Jones Mary S is not matched with Jones Sara A; however, Jones Mary S is matched with Jones Sara M or Jones Missy S). This rule was later revised so that Jones Sara S and Jones Sara M was considered a match. The reasoning was that many Southerners, especially boys, are called by their middle names. Assuming that Mary Sara Jones is called Sara, it would be reasonable that she would code Sara as her first name and correctly code S as her middle initial.

7. Race, sex, or birthdate left blank do not constitute a conflict; "close" birthdays (e.g., 912 and 913) are acceptable as are partial ones which exhibit some consistency (e.g., 912 and bb09 were b represents a blank); month/day reversals are acceptable.

8. Two of the three demographics should match.

9. If first names are different, there must be confirmation data in the middle initial. (Driggers John is not matched with Driggers Cliff but is matched with Driggers Cliff J)

10. First-name last-name reversals are acceptable. (Thomas Nikki S = Nikki Thomas S)

11. Some coding errors are acceptable, particularly if the mistake is obvious by sex associated with first name. (Brown Margaret N = Brown Nancy M even if Nancy has been coded as a male.)